引例：某位同学与一位猎人一起外出打猎，一只野兔从前方窜过．只听一声枪响，野兔应声到下，如果要你推测，这一发命中的子弹是谁打的？你就会想，只发一枪便打中，由于猎人命中的概率一般大于这位同学命中的概率，看来这一枪是猎人射中的．

这个例子所作的推断就体现了极大似然法的基本思想．

一、极大似然思想

一般地说，事件 A 与参数 $\theta \in \Theta$ 有关，θ 取值不同，则 $P(A)$ 也不同．若 A 发生了，则认为此时的 θ 值就是 θ 的估计值．这就是**极大似然思想**．看一例子：**例** 1 、设袋中装有许多黑、白球，不同颜色球的数量比为 3:1，试设计一种方法，估计任取一球为黑球的概率 P ．

分析：易知 P 的值无非是 1/4 或 3/4．为估计 P 的值，现从袋中有放回地任取 3 只球，用 X 表示其中的黑球数，则 $X \sim b(3,P)$ ．按极大似然估计思想，对 P 的取值进行估计．

解：对 P 的不同取值，X 取 $k=0,1,2,3$ 的概率可列表如下:

X	0	1	2	3
$P=1/4$	$27/64$	$27/64$	$9/64$	$1/64$
$P=3/4$	$1/64$	$9/64$	$27/64$	$27/64$

故根据极大似然思想即知：$\hat{P}=\begin{cases} 1/4, k=0,1 \\ 3/4, k=2,3 \end{cases}$.

在上面的例子中，P 是分布中的参数，它只能取两个值：1/4 或 3/4，需要通过抽样来决定分布中参数究竟是 1/4 还是 3/4．在给定了样本观测值后去计算该样本出现的概率，这一概率依赖于 P 的值，为此需要用 1/4、3/4 分别去计算此概率，在相对比较之下，哪个概率大，则 P 就最象那个．

二、似然函数与极大似然估计

1、离散分布场合：

设总体 X 是离散型随机变量，其概率函数为 $p(x;\theta)$ ，其中 θ 是未知参数．设 $X_1,X_2,\cdots,X_n$ 为取自总体 X 的样本． $X_1,X_2,\cdots,X_n$ 的联合概率函数为 $\prod_{i=1}^{n} p(X_i;\theta)$ ，这里，θ 是常量，$X_1,X_2,\cdots,X_n$ 是变量．

若我们已知样本取的值是 $x_1,x_2,\cdots,x_n$ ，则事件 $\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$ 发生的概率为 $\prod_{i=1}^{n} p(x_i;\theta)$ ．这一概率随 θ 的值而变化．从直观上来看，既然样本值 $x_1,x_2,\cdots,x_n$ 出

现了，它们出现的概率相对来说应比较大，应使$\prod_{i=1}^{n} p(x_i;\theta)$取比较大的值．换句话说，$\theta$应使样本值$x_1,x_2,\cdots,x_n$的出现具有最大的概率．将上式看作$\theta$的函数，并用$L(\theta)$表示，就有：

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n} p(x_i;\theta) \qquad （1）$$

称$L(\theta)$为**似然函数**．极大似然估计法就是在参数θ的可能取值范围Θ内，选取使$L(\theta)$达到最大的参数值$\hat{\theta}$，作为参数θ的估计值．即取θ，使

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\hat{\theta})=\max_{\theta\in\Theta} L(x_1,x_2,\cdots,x_n;\theta) \qquad （2）$$

因此，求总体参数θ的极大似然估计值的问题就是求似然函数$L(\theta)$的最大值问题．这可通过解下面的方程$\dfrac{dL(\theta)}{d\theta}=0$　　（3）
来解决．因为$\ln L$是L的增函数，所以$\ln L$与L在θ的同一值处取得最大值．我们称$l(\theta)=\ln L(\theta)$为**对数似然函数**．因此，常将方程（3）写成：$\dfrac{d\ln L(\theta)}{d\theta}=0$
（4）

方程（4）称为**似然方程**．解方程（3）或（4）得到的$\hat{\theta}$就是参数θ的极大似然估计值．

如果方程（4）有唯一解，又能验证它是一个极大值点，则它必是所求的极大似然估计值．有时，直接用（4）式行不通，这时必须回到原始定义（2）进行求解．

2、连续分布场合：

设总体X是连续离散型随机变量，其概率密度函数为$f(x;\theta)$，若取得样本观察值为$x_1,x_2,\cdots,x_n$，则因为随机点$(X_1,X_2,\cdots,X_n)$取值为$(x_1,x_2,\cdots,x_n)$时联合密度函数值为$\prod_{i=1}^{n} f(x_i;\theta)$．所以，按极大似然法，应选择$\theta$的值使此概率达到最大．我们取似然函数为

$L(\theta)=\prod_{i=1}^{n} f(x_i;\theta)$，再按前述方法求参数$\theta$的极大似然估计值．

三、求极大似然估计的方法

1、可通过求导获得极大似然估计：
当函数关于参数可导时，常可通过求导方法来获得似然函数极大值对应的参数值．

例 2、设某工序生产的产品的不合格率为p，抽n个产品作检验，发现有T个不合格，

试求 p 的极大似然估计.

分析：设 X 是抽查一个产品时的不合格品个数，则 X 服从参数为 p 的二点分布 $b(1,p)$．抽查 n 个产品，则得样本 $X_1,X_2,\cdots,X_n$，其观察值为 $x_1,x_2,\cdots,x_n$，假如样本有 T 个不合格，即表示 $x_1,x_2,\cdots,x_n$ 中有 T 个取值为 1， $n-T$ 个取值为 0．按离散分布场合方法，求 p 的极大似然估计.

解：（1）写出似然函数：$L(p)=\prod_{i=1}^{n}p^{x_i}(1-P)^{1-x_i}$

（2）对 $L(p)$ 取对数，得对数似然函数 $l(p)$：

$$l(p)=\sum_{i=1}^{n}[x_i\ln p+(1-x_i)\ln(1-p)]=n\ln(1-p)+\sum_{i=1}^{n}x_i[\ln p-\ln(1-p)]$$

（3）由于 $l(p)$ 对 p 的导数存在，故将 $l(p)$ 对 p 求导，令其为 0，得似然方程：

$$\frac{dl(p)}{dp}=-\frac{n}{1-p}+\sum_{i=1}^{n}x_i(\frac{1}{p}+\frac{1}{1-p})=-\frac{n}{1-p}+\frac{1}{p(1-p)}\sum_{i=1}^{n}x_i=0$$

（4）解似然方程得：$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$

（5）经验证，在 $\hat{p}=\bar{x}$ 时，$\frac{d^2l(p)}{dp^2}<0$，这表明 $\hat{p}=\bar{x}$ 可使似然函数达到最大

（6）上述过程对任一样本观测值都成立，故用样本代替观察值便得 p 的极大似然估计为：$\hat{p}=\overline{X}$

将观察值代入，可得 p 的极大似然估计值为：$\hat{p}=\bar{x}=\frac{T}{n}$，其中 $T=\sum_{i=1}^{n}x_i$．

若总体 X 的分布中含有多个未知参数 $\theta_1,\theta_2,\cdots,\theta_k$ 时，似然函数 L 是这些参数的多元函数 $L(\theta_1,\cdots,\theta_k)$．代替方程（3），我们有方程组 $\frac{\partial(\ln L)}{\partial\theta_i}=0(i=1,2,\cdots,k)$，由这个方程组解得 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$ 分别是参数 $\theta_1,\theta_2,\cdots,\theta_k$ 的极大似然估计值.

例 3、设某机床加工的轴的直径与图纸规定的中心尺寸的偏差服从 $N(\mu,\sigma^2)$，其中 μ,σ^2 未知．为估计 μ,σ^2，从中随机抽取 $n=100$ 根轴，测得其偏差为 $x_1,x_2,\cdots,x_{100}$．试

求 μ,σ^2 的极大似然估计.

分析：显然，该问题是求解含有多个（两个）未知参数的极大似然估计问题．通过建立关于未知参数 μ,σ^2 的似然方程组，从而进行求解.

解：（1）写出似然函数：

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

（2）写出对数似然函数：

$$l(\mu,\sigma^2)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

（3）将 $l(\mu,\sigma^2)$ 分别对 μ、σ^2 求偏导，并令它们都为 0，得似然方程组为：

$$\begin{cases}\dfrac{\partial l(\mu,\sigma^2)}{\partial\mu}=\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2=0\\ \dfrac{\partial l(\mu,\sigma^2)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases}$$

（4）解似然方程组得：

$$\hat{\mu}=\bar{x}，\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

（5）经验证 $\hat{\mu},\hat{\sigma}^2$ 使 $l(\mu,\sigma^2)$ 达到极大，

（6）上述过程对一切样本观察值成立，故用样本代替观察值，便得 μ,σ^2 的极大似然估计分别为：

$$\hat{\mu}=\overline{X}，\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=S_n^2.$$

2、不可通过求导方法获得极大似然估计：

当似然函数的非零区域与未知参数有关时，通常无法通过解似然方程来获得参数的极大似然估计，这时可从定义（2）出发直接求 $L(\theta)$ 的极大值点.

例 4、设总体 X 服从均匀分布 $U(0,\theta)$，从中获得容量为 n 的样本 $X_1,X_2,\cdots,X_n$，其观测值为 $x_1,x_2,\cdots,x_n$，试求 θ 的极大似然估计.

分析：当写出其似然函数 $L(\theta)$ 时，我们会发现 $L(\theta)$ 的非零区域与 θ 有关，因而无法用求导方法来获得 θ 的极大似然估计，从而转向定义（2）直接求 $L(\theta)$ 的极大值.

解：写出似然函数：

$$L(\theta)=\begin{cases}\theta^{-n},0\le x_{(1)}\le x_{(n)}\le\theta\\ 0,\text{其它场合}\end{cases}$$

为使$L(\theta)$达到极大，就必须使θ尽可能小，但是θ不能小于$x_{(n)}$，因而θ取$x_{(n)}$时使$L(\theta)$达到极大，故θ的极大似然估计为：

$$\hat{\theta}=X_{(n)}.$$

进一步，可讨论估计$\hat{\theta}$的无偏性：

由于总体$X \sim U(0,\theta)$，其密度函数与分布函数分别为：$p(x)=\begin{cases}\dfrac{1}{\theta},0<x<\theta\\0,\text{其它}\end{cases}$，$F(x)=\begin{cases}0,x\le 0\\\dfrac{x}{\theta},0<x<\theta\\1,x\ge\theta\end{cases}$，从而$\hat{\theta}=X_{(n)}$的概率密度函数为：$p_{\hat{\theta}}=n[F(y)]^{n-1}p(y)=\dfrac{ny^{n-1}}{\theta^n},0<y<\theta$

$$E(\hat{\theta})=E(X_{(n)})=\int_0^\theta yp_{\hat{\theta}}(y)dy=\int_0^\theta \frac{ny^n}{\theta^n}dy=\frac{n}{n+1}\theta\neq\theta$$

这说明θ的极大似然估计$\hat{\theta}=X_{(n)}$不是θ的无偏估计，但对$\hat{\theta}$作一修正可得θ的无偏估计为：$\hat{\theta}_1=\dfrac{n+1}{n}X_{(n)}$.

通过修正获得未知参数的无偏估计，这是一种常用的方法．在二次世界大战中，从战场上缴获的纳粹德国的枪支上都有一个编号，对最大编号作一修正便获得了德国生产能力的无偏估计．

综上，可得求极大似然估计值的一般步骤．

四、求极大似然估计的一般步骤

1、由总体分布导出样本的联合概率函数（或联合密度）；

2、把样本联合概率函数（或联合密度）中自变量看成已知常数，而把参数θ看作自变量，得到似然函数$L(\theta)$；

3、求似然函数$L(\theta)$的最大值点（常转化为求对数似然函数$l(\theta)$的最大值点）；

4、在最大值点的表达式中，用样本值代入就得参数的极大似然估计值．

五、极大似然估计的不变性

求未知参数θ的某种函数$g(\theta)$的极大似然估计可用极大似然估计的不变原则，证明从略．

定理（不变原则）设$\hat{\theta}$是θ的极大似然估计，$g(\theta)$是θ的连续函数，则$g(\theta)$的极大似然估计为$g(\hat{\theta})$.

例 5、设某元件失效时间服从参数为λ的指数分布，其密度函数为$f(x;\lambda)=\lambda e^{-\lambda x},x\ge 0$，$\lambda$未知．现从中抽取了$n$个元件测得其失效时间为$x_1,x_2,\cdots,x_n$，试求$\lambda$及平均寿命的极大似然估计．

分析：可先求λ的极大似然估计，由于元件的平均寿命即为X的期望值，在指数分布场合，有$E(X)=\dfrac{1}{\lambda}$，它是λ的函数，故可用极大似然估计的不变原则，求其极大似然估计．

解：（1）写出似然函数：$L(\lambda)=\prod_{i=1}^n \lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum_{i=1}^n x_i}$

（2）取对数得对数似然函数：$l(\lambda)=n\ln\lambda-\lambda\sum_{i=1}^n x_i$

（3）将$l(\lambda)$对λ求导得似然方程为：$\dfrac{dl(\lambda)}{d\lambda}=\dfrac{n}{\lambda}-\sum_{i=1}^n x_i=0$

（4）解似然方程得：$\hat{\lambda}=\dfrac{n}{\sum_{i=1}^n x_i}=\dfrac{1}{\bar{x}}$

经验证，$\hat{\lambda}$能使$l(\lambda)$达到最大，由于上述过程对一切样本观察值成立，故λ的极大似然估计为：$\hat{\lambda}=\dfrac{1}{\overline{X}}$；

根据极大似然估计的不变原则，元件的平均寿命的极大似然估计为：$E(X)=\dfrac{1}{\hat{\lambda}}=\overline{X}$.

五、小结

1、极大似然估计的思想；

2、求解未知参数极大似然估计的一般步骤；

3、极大似然估计的不变原则.

五、作业

见参考文献 1 的第 278 页第 4，5，6 页.

参考文献：

1、苏均和主编：概率论与数理统计，上海财经大学出版社．1999 年 1 版.
2、茆诗松等编著：概率论与数理统计，中国统计出版社．1999 年 1 版.
3、魏振军编：概率论与数理统计三十三讲，中国统计出版社．2000 年 1 版.
4、唐生强主编：概率论与数理统计复习指导，科学出版社．1999 年 1 版.

最大似然估计

在本章中我们开始讨论时间序列模型的参数估计方法，其中极大似然估计是一种最为常用的参数估计方法。我们仅仅讨论极大似然估计的原理和似然函数的推导，而对获取极大似然估计的算法不加以详述。

§5.1　引　言

5.1.1　*ARMA* 模型的极大似然估计

假设数据的真实生成过程是一个 $ARMA(p,q)$ 过程，则该过程的数据生成机制为：

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

其中 ε_t 是白噪声序列，满足：

$$E(\varepsilon_s \varepsilon_t) = \begin{cases} \sigma^2, & s = t \\ 0, & s \neq t \end{cases}$$

我们将要讨论如何利用 Y_t 的观测值来估计母体参数：

$$\mathbf{\theta} = (c, \phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q, \sigma^2)$$

我们将要采用的方法是极大似然估计方法，因此需要获得似然函数的表达式。假设获得了 T 个样本 $(y_1, y_2, \cdots, y_T)$，如果能够计算出相应的联合概率密度函数：

$$f_{(Y_1,\cdots,Y_T)}(y_1, y_2, \cdots, y_T; \mathbf{\theta})$$

上述函数可以视为在给定参数下样本发生的概率，因此合理的参数取值是使得上述概率最大，如此参数便称为极大似然估计。这时我们需要极大化上述联合概率密度。

为此，我们假设噪声序列是高斯白噪声序列，即

$\varepsilon_t \sim i.i.d.N(0, \sigma^2)$

虽然这个假设非常强，但是在这样假设下得到的参数估计$\hat{\boldsymbol{\theta}}$，对于非 Gauss 过程来说也是很有意义的。

具体求解极大似然估计的步骤是：一是先求出并计算似然函数，二是求似然函数的最大值。这里涉及到一些代表性的非线性数值优化问题。

§5.2 高斯 $AR(1)$ 过程的似然函数

假设数据生成过程是一个具有高斯白噪声序列的 $AR(1)$ 过程：

$$Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t$$

这时对应的参数向量为：$\boldsymbol{\theta} = (c, \phi, \sigma^2)'$。我们首先寻求联合概率分布函数，也就是这些参数对应的似然函数。

(1) 求上述过程似然函数的代表性过程是利用条件概率密度进行传递，所以需要先求出 Y_1 的概率密度。它的均值和方差为：

$$EY_1 = \frac{c}{1-\phi}，\quad E(Y_1 - \mu)^2 = \frac{\sigma^2}{1-\phi^2}$$

由于它具有正态分析，因此对应的密度函数为：

$$f_{Y_1}(y_1;\boldsymbol{\theta}) = f_{Y_1}(y_1;c,\phi,\sigma^2) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1-\phi^2)}}\exp\left[\frac{-\{y_1-[c/(1-\phi)]\}^2}{2\sigma^2/(1-\phi^2)}\right]$$

(2) 在给定 $Y_1 = y_1$ 的条件下，Y_2 的条件概率分布可以得到：

$$Y_2 | Y_1 = y_1 \sim N((c+\phi y_1), \sigma^2)$$

对应的概率密度函数为：

$$f_{Y_2|Y_1}(y_2 | y_1;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y_2-c-\phi y_1)^2}{2\sigma^2}\right]$$

(3) 类似地，在给定前两个观测值的条件，Y_3 的条件概率密度函数为：

$$f_{Y_3|Y_1,Y_2}(y_3 | y_2, y_1;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y_3-c-\phi y_2)^2}{2\sigma^2}\right]$$

注意到上述条件概率分布中只依赖一阶滞后的条件观测值。

(4) 最后一个样本的条件概率分布为：

$$f_{Y_3|Y_1,Y_2,\cdots,Y_{T-1}}(y_T | y_{T-1},\cdots,y_2,y_1;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y_T-c-\phi y_{T-1})^2}{2\sigma^2}\right]$$

注意到上述条件概率分布中也只依赖一阶滞后的条件观测值。

(5) 根据无条件密度函数与条件密度函数之间的关系，可以得到：

$$f_{Y_T,Y_{T-1},\cdots,Y_1}(y_T, y_{T-1},\cdots,y_2,y_1;\boldsymbol{\theta}) = f_{Y_1}(y_1;\boldsymbol{\theta})\prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t | y_{t-1};\boldsymbol{\theta})$$

经常对上述函数取对数，得到对数似然函数：

$$L(\boldsymbol{\theta}) = \log f_{Y_1}(y_1;\boldsymbol{\theta}) + \sum_{t=2}^{T}\log[f_{Y_t|Y_{t-1}}(y_t | y_{t-1};\boldsymbol{\theta})]$$

(6) 将具体的密度函数代入上式，可以得到 $AR(1)$ 过程的似然函数为：

$$L(\boldsymbol{\theta})=-\frac{1}{2}\log(2\pi)-\frac{1}{2}\log[\sigma^2/(1-\phi^2)]-\frac{\{y_1-[c/(1-\phi)]\}^2}{2\sigma^2/(1-\phi^2)}$$

$$-[(T-1)/2]\log(2\pi)-[(T-1)/2]\log\sigma^2-\sum_{t=2}^{T}\left[\frac{(y_t-c-\phi y_{t-1})^2}{2\sigma^2}\right]$$

可以将上述似然函数表示为更为紧凑的向量和矩阵形式。令均值向量和自协方差为$\boldsymbol{\mu}$ 和 $\boldsymbol{\Omega}$，注意到过程之间具有的自协方差函数表达形式，则有：

$$\boldsymbol{\Omega}=\sigma^2\mathbf{V}\text{，}\quad \mathbf{V}=\frac{1}{1-\phi^2}\begin{bmatrix}1 & \phi & \phi^2 & \cdots & \phi^{T-1}\\ \phi & 1 & \phi & \cdots & \phi^{T-2}\\ \phi^2 & \phi & 1 & \cdots & \phi^{T-3}\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1\end{bmatrix}$$

这样一来，所观测到的样本可以当作多元正态母体 $N(\boldsymbol{\mu},\ \boldsymbol{\Omega})$ 的一个简单抽样，具有的联合概率密度函数为：

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\ =(2\pi)^{-T/2}\,|\boldsymbol{\Omega}^{-1}|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]$$

理论上可以对上述极大似然函数求导数，然后获得参数估计。但是，一般情况下的导数方程是非线性方程，难以获得精确的最大值估计。一种近似的方法是假设第一个观测值是确定性的，然后求解给定 Y_1 时的条件似然函数值，这时的目标函数是：

$$\log f_{Y_T,\cdots,Y_2|Y_1}(y_T,\cdots,y_2\,|\,y_1;\boldsymbol{\theta})=-[(T-1)/2]\log(2\pi)-[(T-1)/2]\log\sigma^2$$
$$-\sum_{t=2}^{T}\left[\frac{(y_t-c-\phi y_{t-1})^2}{2\sigma^2}\right]$$

上式最大值相当于求下式的最小值：

$$\sum_{t=2}^{T}\left[(y_t-c-\phi y_{t-1})^2\right]$$

上式的最小值就是线性回归的最小二乘估计，满足方程：

$$\begin{bmatrix}\hat{c}\\ \hat{\phi}\end{bmatrix}=\begin{bmatrix}T-1 & \sum_{t=2}^{T}y_{t-1}\\ \sum_{t=2}^{T}y_{t-1} & \sum_{t=2}^{T}y_{t-1}^2\end{bmatrix}^{-1}\begin{bmatrix}\sum_{t=2}^{T}y_t\\ \sum_{t=2}^{T}y_{t-1}y_t\end{bmatrix}$$

类似地，噪声的方差为：

$$\hat{\sigma}^2=\frac{1}{T-1}\sum_{t=2}^{T}\left[(y_t-\hat{c}-\hat{\phi}y_{t-1})^2\right]$$

当样本容量足够大时，可以证明上述近似或者条件极大似然估计具有与精确极大似然估计一致的极限分布。

§5.3 高斯 $AR(p)$ 过程的似然函数

对于一般的高阶自回归过程：

$$Y_t=c+\phi_1Y_{t-1}+\phi_2Y_{t-2}+\cdots+\phi_pY_{t-p}+\varepsilon_t\text{，}\quad \varepsilon_t\sim i.i.d\,N(0,\sigma^2)$$

此时所要估计的总体参数向量是：$\boldsymbol{\theta}=(c,\phi_1,\phi_2,\cdots,\phi_p,\sigma^2)$。

(1) 似然函数的估值 Evaluating the Likelihood Function

假设我们获得了 T 个来自 $AR(p)$ 过程的样本，假设前 p 个样本表示为

$\mathbf{y}_p = (y_1, y_2, \cdots, y_p)'$

可以将这个向量当作 p 维 Gauss 变量的一个样本。这个向量的均值表示为 $\boldsymbol{\mu}_p$，它的每个分量都是：

$\mu = c/(1-\phi_1-\cdots-\phi_p)$

假设 $\sigma^2\mathbf{V}_p$ 是 $(\mathbf{Y}_1,\cdots,\mathbf{Y}_p)$ 的协方差矩阵，则有：

$$\sigma^2\mathbf{V}_p = \begin{bmatrix} E(Y_1-\mu)^2 & E[(Y_1-\mu)(Y_2-\mu)] & \cdots & E[(Y_1-\mu)(Y_p-\mu)] \\ E[(Y_2-\mu)(Y_1-\mu)] & E(Y_2-\mu)^2 & \cdots & E[(Y_2-\mu)(Y_p-\mu)] \\ \vdots & \vdots & \cdots & \vdots \\ E[(Y_p-\mu)(Y_1-\mu)] & E[(Y_p-\mu)(Y_2-\mu)] & \cdots & E(Y_p-\mu)^2 \end{bmatrix}$$

对于一阶自回归过程而言（$p=1$），上述矩阵是一个标量，$\mathbf{V}_p = 1/(1-\phi^2)$；对于 p 阶自回归而言：

$$\sigma^2\mathbf{V}_p = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p=1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p=2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p=3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_{p=1} & \gamma_{p=2} & \gamma_{p=3} & \cdots & \gamma_0 \end{bmatrix}$$

这里 γ_j 是 $AR(p)$ 过程的第 j 个自协方差，可以按照以前的介绍公式计算。

由于自回归过程的条件相依性具有截断性质，因此我们将样本分为 p 个一组，样本中前 p 个观测值的联合概率分布为 $N(\boldsymbol{\mu}_p, \sigma^2\mathbf{V}_p)$，密度为：

$$\begin{aligned} & f_{Y_p,Y_{p-1},\cdots,Y_1}(Y_p, Y_{p-1},\cdots,Y_1;\boldsymbol{\theta}) \\ &= (2\pi)^{-p/2}|\sigma^{-2}\mathbf{V}_p^{-1}|^{1/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}_p-\boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p-\boldsymbol{\mu}_p)\right] \\ &= (2\pi)^{-p/2}(\sigma^{-2})^{p/2}|\mathbf{V}_p^{-1}|^{1/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}_p-\boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p-\boldsymbol{\mu}_p)\right] \end{aligned}$$

对于样本中剩余的观测值 $(y_{p+1}, y_{p+2},\cdots,y_T)$，我们可以使用推断误差分解(prediction error decomposition)，将前 $t-1$ 个观测值作为条件，则第 t 个观测值的条件分布为 Gauss 分布，且均值和方差分别为：

$c+\phi_1 y_{t-1}+\phi_2 y_{t-2}+\cdots+\phi_p y_{t-p}$，$\sigma^2$

只有 p 个最近的观测值与这个分布有关，因此，对于 $t>p$，则有：

$$\begin{aligned} & f_{Y_t|Y_{t-1},Y_{t-2},\cdots,Y_1}(y_t|y_{t-1},y_{t-2},\cdots,y_1;\boldsymbol{\theta}) = f_{Y_t|Y_p,Y_{p-1},\cdots,Y_{t-p}}(y_t|y_{t-1},y_{t-2},\cdots,y_{t-p};\boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y_t-c-\phi_1 y_{t-1}-\phi_2 y_{t-2}-\cdots-\phi_p y_{t-p})^2\right] \end{aligned}$$

因此，整个样本的似然函数为：

$$\begin{aligned} & f_{Y_T\,Y_{T-1},,\cdots,Y_1}(y_T,\, y_{T-1}, y_{T-2},\cdots,y_1;\boldsymbol{\theta}) = \\ & \quad f_{Y_p\,Y_{p-1},,\cdots,Y_1}(y_p,\, y_{p-1}, y_{p-2},\cdots,y_1;\boldsymbol{\theta}) \times \prod_{t=p+1}^{T} {}_{Y_t|Y_{t-1},Y_{p-2},\cdots,Y_{1t-p}}(y_t|y_{t-1},y_{t-2},\cdots,y_{t-p};\boldsymbol{\theta}) \end{aligned}$$

则对数似然函数形式为：

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \log f_{Y_t|Y_{t-1},Y_{p-2},\cdots,Y_{11}}(y_t|y_{t-1},y_{t-2},\cdots,y_1;\boldsymbol{\theta}) \\
&= -\frac{p}{2}\log(2\pi)-\frac{p}{2}\log(\sigma^2)+\frac{1}{2}\log|\mathbf{V}_p^{-1}|-\frac{1}{2\sigma^2}(\mathbf{y}_p-\boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p-\boldsymbol{\mu}_p) \\
&\quad -\frac{T-p}{2}\log(2\pi)-\frac{T-p}{2}\log(\sigma^2)-\sum_{t=p+1}^{T}\frac{(y_t-c-\phi_1 y_{t-1}-\phi_2 y_{t-2}-\cdots-\phi_p y_{t-p})^2}{2\sigma^2} \\
&= -\frac{T}{2}\log(2\pi)-\frac{T}{2}\log(\sigma^2)+\frac{1}{2}\log|\mathbf{V}_p^{-1}|-\frac{1}{2\sigma^2}(\mathbf{y}_p-\boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p-\boldsymbol{\mu}_p) \\
&\quad -\sum_{t=p+1}^{T}\frac{(y_t-c-\phi_1 y_{t-1}-\phi_2 y_{t-2}-\cdots-\phi_p y_{t-p})^2}{2\sigma^2}
\end{aligned}
$$

为了获得上述似然函数值，我们需要获得逆矩阵$\mathbf{V}_p^{-1}$，为此我们有下述命题：

命题 5.1 利用$v^{ij}(p)$表示矩阵$\mathbf{V}_p^{-1}$的第(i,j)位置的元素，则对任意$1\le i\le j\le p$，有：

$$
v^{ij}(p)=\left[\sum_{k=0}^{i-1}\phi_k\phi_{k+j-i}-\sum_{k=p+1-j}^{p+i-j}\phi_k\phi_{k+j-i}\right]
$$

这里$\phi_0\equiv -1$。

证明：略。 **End**

因为$\mathbf{V}_p^{-1}$是对角矩阵，因此也可以得到$i>j$时的元素$v^{ij}(p)$。

例如，对$AR(1)$过程而言，$\mathbf{V}_p^{-1}$是一个标量，取$i=j=p=1$，得到：

$$
\mathbf{V}_1^{-1}=v^{11}(1)=\left[\sum_{k=0}^{0}\phi_k\phi_k-\sum_{k=1}^{1}\phi_k\phi_k\right]=(\phi_0^2-\phi_1^2)=(1-\phi^2)
$$

因此有：

$$
\sigma^2\mathbf{V}_1=\frac{\sigma^2}{1-\phi^2}
$$

因此命题 5.1 确实可以重新得到$AR(1)$过程的方差表达式。

对于$p=2$的情形，利用命题 5.1 可以得到：

$$
\mathbf{V}_2^{-1}=\begin{bmatrix}1-\phi_2^2 & -(\phi_1+\phi_1\phi_2)\\ -(\phi_1+\phi_1\phi_2) & 1-\phi_2^2\end{bmatrix}
$$

可以计算行列式值为：

$$
|\mathbf{V}_2^{-1}|=\left|(1+\phi_2)\begin{bmatrix}1-\phi_2 & -\phi_1\\ -\phi_1 & 1-\phi_2\end{bmatrix}\right|=(1+\phi_2)^2[(1-\phi_2)^2-\phi_1^2]
$$

并且有：

$$
\begin{aligned}
&(\mathbf{y}_2-\boldsymbol{\mu}_2)'\mathbf{V}_2^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2) \\
&\quad=[(y_1-\mu),(y_2-\mu)](1+\phi_2)\begin{bmatrix}1-\phi_2 & -\phi_1\\ -\phi_1 & 1-\phi_2\end{bmatrix}\begin{bmatrix}(y_1-\mu)\\ (y_2-\mu)\end{bmatrix} \\
&\quad=(1+\phi_2)\times\{(1-\phi_2)(y_1-\mu)^2-2\phi_1(y_1-\mu)(y_2-\mu)+(1-\phi_2)(y_2-\mu)^2\}
\end{aligned}
$$

因此，对于 Gauss 条件下的$AR(2)$过程，确切的似然函数为：

$$L(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log\{(1+\phi_2)^2[(1-\phi_2)^2 - \phi_1^2]\}$$
$$-\left\{\frac{1+\phi_2}{2\sigma^2}\right\} \times \{(1-\phi_2)(y_1-\mu)^2 - 2\phi_1(y_1-\mu)(y_2-\mu) + (1-\phi_2)(y_2-\mu)^2\}$$
$$-\sum_{t=3}^{T}\frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2})^2}{2\sigma^2}$$

这里：$\mu = c/(1-\phi_1-\phi_2)$

(2) 条件极大似然估计 Conditional Maximum Likelihood Estimates

由于目标函数形式比较复杂，因此对 $AR(p)$ 过程的确切极大化必须使用数值算法。与此对应，以前 p 个样本为条件的对数似然函数具有下述简单形式：

$$\log f_{Y_T, Y_{T-1}, \cdots, Y_{p+1}|Y_{t-1}, Y_{p-2}, \cdots, Y_1}(y_T, y_{T-1}, \cdots, y_{p+1} | y_p, y_{t-2}, \cdots, y_1; \boldsymbol{\theta})$$
$$= -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T}\frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}$$

注意到，极大化上式的参数 $(c, \phi_1, \phi_2, \cdots, \phi_p)$ 与极小化下式的参数选择是一致的：

$$\sum_{t=p+1}^{T}(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2$$

因此这些参数 $(c, \phi_1, \phi_2, \cdots, \phi_p)$ 的条件极大似然估计是 y_t 基于常数和自身滞后值的普通最小二乘回归估计，σ^2 的条件极大似然估计是这个回归方程平方残差的平均值：

$$\hat{\sigma}^2 = \frac{1}{T-p}\sum_{t=p+1}^{T}(y_t - \hat{c} - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2} - \cdots - \hat{\phi}_p y_{t-p})^2$$

显然上述条件似然函数与确切似然函数相比，缺少了初始样本的母体分布，这样就降低了样本发生的似然性，这就是条件似然函数与确切似然函数的差异。

类似地，确切的极大似然估计和条件极大似然估计能够得到相同的大样本分布。

(3) 非 Gauss 分布时间序列的极大似然估计 Maximum Likelihood Estimation for Non-Gaussian Time Series

根据线性回归模型的性质，如果假设随机过程关于二阶矩是遍历的，则我们知道普通最小二乘估计也是下面线性投影系数的一致估计(consistent estimate)：

$$\hat{E}(Y_t | Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p})$$

同时这个 OLS 估计也使得 Gauss 条件似然函数达到最大。因此，即使一个过程不是 Gauss 过程，但是我们错误地将它当作 Gauss 过程，并且极大化它的似然函数，则得到参数估计 $(\hat{c}, \hat{\phi}_1, \hat{\phi}_2, \cdots, \hat{\phi}_p)$ 仍然是母体参数的一致估计。

一个从错误指定的似然函数中获得的极大似然估计(例如，当真实过程不是 Gauss 过程，而从 Gauss 过程假设下获得的 MLE)被称为“伪极大似然估计”(quasi-maximum likelihood estimate)。但是，有些时候，例如上述提到的情形，“伪极大似然估计”能够为人们感兴趣的参数提供一致估计。但是，如果真实数据不是服从 Gauss 过程的，在 Gauss 假设下获得参数估计的标准差则是不正确的，它不具有估计的一致性。

另外，如果原始数据不是服从 Gauss 过程的，有时通过简单的变化，例如取对数等，可以获得 Gauss 分布过程。对于一个正的随机变量 Y_t，Box and Cox (1964)提出了一类比较一般的函数变换：

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{Y_t^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log Y_t, & \lambda = 0 \end{cases}$$

一种方法是，选取特殊的λ，在$Y_t^{(\lambda)}$是高斯*ARMA*过程的假设下极大化$Y_t^{(\lambda)}$的似然函数，这时能够使得似然函数取最大值的λ被认为是最合适的变换选择。但是，也有一些研究者对这种方法在实用中的效果提出了质疑。

§5.4　高斯$MA(1)$过程的似然函数

利用类似自回归过程似然函数的推导方法，我们开始类似地讨论移动平均过程的极大似然估计问题。

5.4.1　条件似然函数 Conditional Likelihood Function

如果基于Y的初始值，则自回归过程的条件似然函数计算是相当简单的。类似地，如果利用初始的ε作为条件，移动平均过程的似然函数计算也是比较简单的。

考虑一个高斯$MA(1)$过程：

$$Y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1}, \quad \varepsilon_t \sim i.i.d\, N(0, \sigma^2)$$

利用参数向量$\boldsymbol{\theta} = (\mu, \theta, \sigma^2)'$表示母体参数，如果$\varepsilon_{t-1}$的数值是确定可知的，则：

$$Y_t \mid \varepsilon_{t-1} \sim N[(\mu + \theta\varepsilon_{t-1}), \sigma^2]$$

表示成为密度函数情形为：

$$f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_t - \mu - \theta\varepsilon_{t-1})^2\right]$$

假设我们确定性地已知$\varepsilon_0 = 0$，并且获得观测值y_1，则下一个阶段的误差也是确定性已知的，可以确定为：

$$\varepsilon_1 = y_1 - \mu$$

利用上述条件概率分布公式得到：

$$f_{Y_2|Y_1,\varepsilon_0=0}(y_2|y_1, \varepsilon_0, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_2 - \mu - \theta\varepsilon_1)^2\right]$$

由于ε_1是确定性可知的，ε_2可以按照下式计算：

$$\varepsilon_2 = y_2 - \mu - \theta\varepsilon_1$$

因此，已知初值$\varepsilon_0 = 0$，可以利用叠代方式从样本$\{y_1, y_2, \cdots, y_T\}$中获得误差序列$\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_T\}$：

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}, \quad t = 1, 2, \ldots, T$$

则条件概率分布为：

$$f_{Y_t|Y_{t-1},Y_{t-2},\cdots,Y_1\,\varepsilon_0=0}(y_t \mid y_{t-1}, y_{t-2}, \cdots, y_1,\, \varepsilon_0 = 0; \boldsymbol{\theta}) = f_{Y_t|\varepsilon_{t-1}=0}(y_t \mid \varepsilon_{t-1}; \boldsymbol{\theta})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\varepsilon_t^2}{2\sigma^2}\right]$$

则联合概率分布可以通过单独的条件概率密度的乘积得到：

$$f_{Y_t,Y_{t-1},Y_{t-2},\cdots,Y_1|\varepsilon_0=0}(y_T, y_{T-1}, y_{T-2}, \cdots, y_1|\varepsilon_0 = 0; \boldsymbol{\theta})$$

$$= f_{Y_1|\varepsilon_0=0}(y_1 \mid \varepsilon_0; \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t|Y_{t-1},Y_{t2},\cdots,Y_1,\varepsilon_0=0}(y_t \mid y_{t-1}, y_{t-2}, \cdots, y_1, \varepsilon_0 = 0; \boldsymbol{\theta})$$

则条件对数似然函数为：

$$L(\boldsymbol{\theta}) = \log f_{Y_t, Y_{t-1}, Y_{t-2}, \cdots, Y_1 | \varepsilon_0 = 0}(y_T, y_{T-1}, y_{T-2}, \cdots, y_1 | \varepsilon_0 = 0; \boldsymbol{\theta})$$

$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}$$

虽然上述函数比较简单，但由于是参数的非线性函数，无法直接利用解析方法获得参数的极大似然估计。因此，即使是 1 阶移动平均过程，条件极大似然估计也需要利用数值优化方法求解。

从任意初始误差ε_0开始叠代，可以得到：

$$\varepsilon_t = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - \cdots + (-1)^{t-1}\theta^{t-1}(y_1 - \mu) + (-1)^t\theta^t\varepsilon_0$$

如果$|\theta|$显著地比 1 小，则附加初始条件$\varepsilon_0 = 0$带来的影响将快速地消失，则当样本数量比较大的时候，条件似然函数是无条件似然函数比较好的近似。相反地，如果$|\theta|>1$，则附加$\varepsilon_0 = 0$带来的结果将随着时间而累积，在这种情形下条件概率方式不是很好的近似方法。这时如果数值方法模拟的结果是$|\hat{\theta}|>1$，则需要放弃这样的估计结果。这是需要利用$\hat{\theta}$的倒数作为初值重新进行条件似然估计的数值计算。

5.4.2　确切似然函数 Exact Likelihood Function

对于高斯$MA(1)$过程，有两种比较合适的算法来计算确切似然函数。一种是使用后面介绍的卡尔曼滤波方法，另一种是方差或协方差矩阵的三角因子分解。下面我们开始介绍矩阵的三角因子分解方法。

假设样本观测值可以利用$T \times 1$矩阵表示：$\mathbf{y} \equiv (y_1, y_2, \cdots, y_T)'$，这个向量具有均值为$\boldsymbol{\mu} = (\mu, \mu, \cdots, \mu)'$和$T \times T$的方差—协方差矩阵$\boldsymbol{\Omega}$：

$$\boldsymbol{\Omega} = E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})']$$

从高斯$MA(1)$过程中获取样本的方差—协方差矩阵为：

$$\boldsymbol{\Omega} = \sigma^2\begin{bmatrix} (1+\theta^2) & \theta & 0 & \cdots & 0 \\ \theta & (1+\theta^2) & \theta & \cdots & 0 \\ 0 & \theta & (1+\theta^2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & (1+\theta^2) \end{bmatrix}$$

则联合分布表示的似然函数是：

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) = (2\pi)^{-T/2}\,|\boldsymbol{\Omega}^{-1}|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]$$

似然性的推断误差分解可以从下面矩阵$\boldsymbol{\Omega}$的三角因子化中获得：

$$\boldsymbol{\Omega} = \mathbf{ADA}'$$

这里的矩阵$\mathbf{A}$是下面形式的下三角矩阵：

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \dfrac{\theta}{1+\theta^2} & 1 & 0 & \cdots & 0 & 0 \\ 0 & \dfrac{\theta(1+\theta^2)}{1+\theta^2+\theta^4} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \dfrac{\theta(1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)})}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} & 1 \end{bmatrix}$$

矩阵 $\mathbf{D}$ 是下述形式的对角矩阵

$$\mathbf{D}=\begin{bmatrix} 1+\theta^2 & 0 & 0 & \cdots & 0 \\ 0 & \dfrac{1+\theta^2+\theta^4}{1+\theta^2} & & \cdots & 0 \\ 0 & 0 & \dfrac{1+\theta^2+\theta^4+\theta^6}{1+\theta^2+\theta^4} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \dfrac{1+\theta^2+\theta^4+\cdots+\theta^{2n}}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} \end{bmatrix}$$

将一个对角矩阵表示成为下三角矩阵的对角化，称为一个矩阵的三角因子化。将这样的矩阵分解带入到似然函数中，得到：

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})=(2\pi)^{-T/2}\,|\mathbf{ADA}'|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'[\mathbf{A}']^{-1}\mathbf{D}^{-1}\mathbf{A}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]$$

注意到矩阵 $\mathbf{A}$ 是主对角线都是 1 的下三角矩阵，因此它的行列式为 $|\mathbf{A}|=1$，这时有：

$$|\mathbf{ADA}'|=|\mathbf{A}||\mathbf{D}||\mathbf{A}'|=|\mathbf{D}|$$

进一步定义：

$$\tilde{\mathbf{y}}\equiv\mathbf{A}^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

则似然函数可以表示成为：

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})=(2\pi)^{-T/2}\,|\mathbf{D}|^{-1/2}\exp\left[-\frac{1}{2}\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}\right]$$

注意到变换形式也可以表示为：

$$\mathbf{A}\tilde{\mathbf{y}}=\mathbf{y}-\boldsymbol{\mu}$$

这个系统的第一行表明 $\tilde{y}_1=y_1-\mu$，而第 t 行意味着：

$$\tilde{y}_t=y_t-\mu-\frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(t-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(t-1)}}\tilde{y}_{t-1}$$

因此向量 $\tilde{\mathbf{y}}$ 可以从 $\tilde{y}_1=y_1-\mu$ 开始，利用上式进行叠代得到，$t=2,3,\cdots,T$。变量 $\tilde{y}_t$ 可以解释为 y_t 基于常数和 $y_{t-1},y_{t-2},\cdots,y_1$ 的线性投影，而矩阵 $\mathbf{D}$ 的第 t 个对角元素是这个线性投影的 MSE：

$$d_{tt}=E(\tilde{Y}_t^2)=\sigma^2\frac{1+\theta^2+\theta^4+\cdots+\theta^{2t}}{1+\theta^2+\theta^4+\cdots+\theta^{2(t-1)}}$$

由于矩阵 $\mathbf{D}$ 是对角矩阵，其行列式是对角元素的乘积：

$$|\mathbf{D}|=\prod_{t=1}^{T}d_{tt}$$

而矩阵 $\mathbf{D}^{-1}$ 可以利用矩阵 $\mathbf{D}$ 的主对角线元素取倒数得到。因此：

$$\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}=\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}$$

这样一来，似然函数可以简化为：

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})=(2\pi)^{-T/2}\left[\prod_{t=1}^{T}d_{tt}\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}\right]$$

因此，高斯 $MA(1)$ 过程的确切对数似然函数为：

$$L(\boldsymbol{\theta})=\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})=-\frac{T}{2}\log(2\pi)-\frac{1}{2}\sum_{t=1}^{T}\log(d_{tt})-\frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}$$

如果给定参数的初始数值，可以利用叠代方法计算上述似然函数。这样的计算可以对尝试参数的优劣进行比较。

与条件似然函数不同的是，上述似然函数的表达式对任何参数 θ 都是有效的，而不取决于是否该过程具有可逆 $MA(1)$ 表示。

§ 5.5 高斯 $MA(q)$ 过程的似然函数

在讨论了一阶移动平均过程似然函数的基础上，我们继续讨论 $MA(q)$ 过程的极大似然估计问题。

5.5.1 高斯 $MA(q)$ 过程的条件似然函数 Conditional Likelihood Function

考虑一个高斯 $MA(q)$ 过程：

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim i.i.d\, N(0, \sigma^2)$$

此时一个简单方法是考虑给定前 q 个 ε 初值的条件似然函数，假设初值为：

$$\varepsilon_0 = \varepsilon_{-1} = \varepsilon_{-2} = \cdots = \varepsilon_{-q+1} = 0$$

利用这些初值，我们对时刻 $t = 1, 2, \ldots, T$，利用下式进行叠代：

$$\varepsilon_t = y_t - \mu - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

利用向量 $\boldsymbol{\varepsilon}_0$ 表示 $q \times 1$ 向量 $(\varepsilon_0, \varepsilon_{-1}, \cdots, \varepsilon_{-q+1})'$，则条件对数似然函数为：

$$L(\boldsymbol{\theta}) = \log f_{Y_t, Y_{t-1}, Y_{t-2}, \cdots, Y_1 | \boldsymbol{\varepsilon}_0 = 0}(y_T, y_{T-1}, y_{T-2}, \cdots, y_1 | \boldsymbol{\varepsilon}_0 = 0; \boldsymbol{\theta})$$

$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}$$

这里的参数向量为：$\boldsymbol{\theta} = (\mu, \theta_1, \theta_2, \cdots, \theta_q)'$。与 1 阶情形类似，这里要求下述特征方程的根落在单位圆外面：

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

上述条件是要求移动平均过程是可逆的，这时条件似然函数的参数估计对误差初值的选取没有严重的依赖性。

5.5.2 高斯 $MA(q)$ 过程的确切似然函数 Exact Likelihood Function

由于高斯过程的联合分布是多元正态概率分布，因此高斯 $MA(q)$ 过程的确切似然函数可以表示为：

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2} |\boldsymbol{\Omega}^{-1}|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right]$$

这里的样本为 $\mathbf{y} \equiv (y_1, y_2, \cdots, y_T)'$，均值为 $\boldsymbol{\mu} = (\mu, \mu, \cdots, \mu)'$，$T \times T$ 的方差—协方差矩阵为：

$$\boldsymbol{\Omega} = \begin{bmatrix} & & & & & \ddots \\ & & & & \gamma_q & \\ & & & \ddots & & \\ & & \gamma_0 & & & \\ & \ddots & & & & \\ \gamma_q & & & & & \\ \ddots & & & & & \end{bmatrix}$$

上述矩阵中的元素为：

$$\mathbf{\Omega}(i,j)=\begin{cases}\gamma_{|i-j|}, & |i-j|\le q\\ 0, & |i-j|>q\end{cases}$$

这里γ_k是$MA(q)$过程的k阶自协方差函数：

$$\gamma_k=\begin{cases}\sigma^2(\theta_k+\theta_{k+1}\theta_1+\theta_{k+2}\theta_2+\cdots+\theta_q\theta_{q-k}), & k=0,1,2,\cdots,q\\ 0, & k>q\end{cases}，\quad \theta_0\equiv 1$$

与一阶情形是一样的，确切似然函数可以通过卡尔曼滤波或者矩阵$\mathbf{\Omega}$的三角因子化获得。仍然考虑三角因子化，这时矩阵$\mathbf{\Omega}$可以分解为：

$$\mathbf{\Omega}=\mathbf{ADA}'$$

这里的矩阵$\mathbf{A}$是下三角矩阵，矩阵$\mathbf{D}$是对角矩阵，它们具有前面介绍预测时的表达式类似。进一步分析可以发现，由于矩阵$\mathbf{\Omega}$具有的特殊谱结构，矩阵$\mathbf{A}$具有性质：对于$i>q+j$，$a_{ij}=0$。对数给定的数值参数，计算机程序可以十分容易地获得这些矩阵的数值。

将这样的矩阵分解带入到$MA(q)$过程的似然函数中，得到：

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})=(2\pi)^{-T/2}|\mathbf{D}|^{-1/2}\exp\left[-\frac{1}{2}\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}\right]$$

这里：$\mathbf{A}\tilde{\mathbf{y}}=\mathbf{y}-\boldsymbol{\mu}$，$\tilde{\mathbf{y}}$的元素可以叠代计算如下：

$$\tilde{y}_1=y_1-\mu,$$
$$\tilde{y}_2=(y_2-\mu)-a_{21}\tilde{y}_1,$$
$$\tilde{y}_3=(y_3-\mu)-a_{32}\tilde{y}_2-a_{31}\tilde{y}_1,$$
$$\vdots$$
$$\tilde{y}_t=(y_t-\mu)-a_{t,t-1}\tilde{y}_{t-1}-a_{t,t-2}\tilde{y}_{t-2}-\cdots-a_{t,t-q}\tilde{y}_{t-q}$$

因此，高斯$MA(q)$过程的确切对数似然函数为：

$$L(\boldsymbol{\theta})=\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})=-\frac{T}{2}\log(2\pi)-\frac{1}{2}\sum_{t=1}^{T}\log(d_{tt})-\frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}$$

给定参数数值，可以计算上述似然函数的函数值。但是，给定样本，计算上述函数的极大值的非线性优化过程却是比较复杂的。

§5.6　高斯$ARMA(p,q)$过程的似然函数

将自回归过程和移动平均过程的方法结合起来，就可以讨论移动平均过程的条件似然函数和确切似然函数问题。

5.6.1　条件似然函数 Conditional Likelihood Function

考虑一个高斯$ARMA(p,q)$过程：

$$Y_t=c+\phi_1 Y_{t-1}+\phi_2 Y_{t-2}+\cdots+\phi_p Y_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q},$$
$$\varepsilon_t\sim i.i.d\,N(0,\sigma^2)$$

我们的目标是估计母体参数向量$\boldsymbol{\theta}=(c,\phi_1,\phi_2,\cdots,\phi_p\theta_1,\theta_2,\cdots,\theta_q,\sigma^2)'$。

此时，对自回归部分条件似然函数的近似依赖于初始的y，对移动平均部分条件似然函数的近似依赖于初始的ε。对移动平均过程$ARMA(p,q)$的似然函数共同的近似即依赖初始的y，也依赖初始的ε。

将初始的$\mathbf{y}_0\equiv(y_0,y_{-1},\cdots,y_{-p+1})'$和$\boldsymbol{\varepsilon}_0\equiv(\varepsilon_0,\varepsilon_{-1},\cdots,\varepsilon_{-q+1})'$当作给定的，则误差序列$\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_T\}$可以通过$\{y_1,y_2,\cdots,y_T\}$，利用下式叠代获得：

$$\varepsilon_t = y_t - c - \phi_1 y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}\,,$$

$t = 1, 2, \ldots, T$

因此条件似然函数是：

$$L(\boldsymbol{\theta}) = \log f_{Y_t, Y_{t-1}, Y_{t-2}, \cdots, Y_1 | \mathbf{Y}_0, \boldsymbol{\varepsilon}_0}(y_T, y_{T-1}, y_{T-2}, \cdots, y_1 | \mathbf{Y}_0, \boldsymbol{\varepsilon}_0; \boldsymbol{\theta})$$

$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}$$

一种选择是将 y 和 ε 的初值设定为它们的预期值，即：

$$y_s = \frac{c}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}\,,\quad s = 0, -1, -2, \ldots, -p+1$$

$\varepsilon_s = 0$， $s = 0, -1, -2, \ldots, -q+1$

然后对 $t = 1, 2, \ldots, T$ ，进行上述的叠代计算。

另外，Box and Jenkins (1976) 建议设定 ε 为零，但将 y 设定为它们的真实数值。因此，对上式的叠代可以从时点 $t = p+1$ 开始，将 $\{y_1, y_2, \cdots, y_p\}$ 选取为真实的样本观测值，选取初值：

$$\varepsilon_p = \varepsilon_{p-1} = \cdots = \varepsilon_{p-q+1} = 0$$

则对数条件似然函数可以计算为：

$$L(\boldsymbol{\theta}) = \log f(y_T, y_{T-1}, y_{T-2}, \cdots, y_{p+1} | y_p, y_{p-1}, \cdots, y_1; \varepsilon_p = 0, \varepsilon_{p-1} = 0, \cdots, \varepsilon_{p-q+1} = 0)$$

$$= -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}$$

如同移动平均过程的情形，这种近似只有当下述特征方程：

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

的特征根全部落在单位圆外面才能应用。